Measuring the relevance of factors in the occurrences of events

Vito Fragnelli*, Josep Freixas†, Montserrat Pons†, and Lluís Sanmiquel‡

April 6, 2014

Abstract

A new way to compare the relevance of the different factors intervening in the occurrences of an event is presented and developed in this paper. The idea behind the method comes from cooperative game theory but the focus is slightly different because factors are not necessarily rational decision-makers and because the only data available are obtained by repetition of the event. The concept of relevance measure for a factor in a set of data is introduced, some significant examples are given and the main properties of relevance measures are defined and studied. One of these measures, the fair measure, is proved to have interesting properties which characterize it. Two real world situations, one about traffic accidents and the other one about mining accidents, both of them with real data, are used to show the use of relevance measures to compare factors in each one of these events.

Keywords: Relevance measures; Decision-Making; Budget distribution.

1 Introduction

Game theory is the study of mathematical models of conflict and cooperation between rational decision-makers. In game theory, a cooperative game is a game where groups of players (coalitions) may enforce cooperative behavior; hence the game is a competition between coalitions of players, rather than between individual players. Several types of solution have been analyzed in depth; among them excel the core ([14] and [15]) as a set solution concept, and the Shapley value [32] as a one-point solution concept. One-point solutions quantify how much players should receive in dividing the spoil. For problems involving revenues (or costs) a one-point solution assigns to each player the total amount he/she should receive (or pay) for cooperating in the game. More particularly, one-point solutions like the power indices are used for studying the strategic importance that players have in a committee, such as a Parliament or a stockholder society (see [1], [13], [25], [34], [35]). Although game theory is thought for players being decision-makers, it is also possible to apply some of its techniques when players are not rational decision-makers. Examples can be found, among others, in medicine ([11], [24], [26]) or in engineering ([2], [3], [9], [12], [19]).

*Dipartimento di Scienze e Innovazione Tecnologica, Università del Piemonte Orientale.
E-mail: *vito.fragnelli@mfn.unipmn.it.*

†Departament de Matemàtica Aplicada 3 i Escola Politècnica Superior d'Enginyeria de Manresa, Universitat Politècnica de Catalunya.
E-mails: *[josep.freixas,montserrat.pons]@upc.edu.*

‡Departament d'Enginyeria Minera i Recursos Naturals i Escola Politècnica Superior d'Enginyeria de Manresa, Universitat Politècnica de Catalunya.
E-mail: *sanmi@emrn.upc.edu.*

The Shapley value [32] assigns to each individual the expected payoff obtained as a result of the cooperation. The approach is based on a list of reasonable properties (axioms) that a value should fulfill, and there is a unique value that meets all these properties, understanding that the list of properties is independent, i.e., minimal. In [27] one can find a good survey on the transversal use of this value, since it applies to many different fields like medicine, reliability or telecommunications, in which the players are, respectively, genes, components or antennas. In particular, the Shapley value has been used in multiple regression analysis to estimate the relative impact of the different predictors (see [21]).

The possibilities of exploiting in different ways the Shapley value can be derived by the several axiomatizations that are available, each based on a different set of properties; it is also possible to consider that some interesting methods for sharing a profit or a cost, or for evaluating the role of the agents in a complex system, based on assumptions which do not use requirements of cooperative games, often result in the Shapley value of a suitable cooperative game. For instance, in [22] the authors say that "*The Shapley value presents each participant's input in the game over all possible combinations of players - in the case of regression, these are the independent variables used in fitting the data by the model. The predictors can be interpreted as players - representative of the real players, respondents - whose opinions constitute the observations. The respondents - via the predictors - define the results of the arbitration, or contribution of the regressors in the quality of the model*".We can mention also the Baker-Thompson rule that coincides with the Shapley value of the airport game (see [23] and [10]) or the fuzzy Choquet integral in decision making and the "lower entropy" approach in information theory that coincide with the Shapley value of a suitable game. The Shapley value was used in marketing decision analysis for estimating the influence of key drivers on customer satisfaction and the strength of the items in covering the maximum number of consumers (see [7], [6], and [20]).

In this paper we present a different and novel use of cooperative games whose aim is to obtain a direct measure of the importance of the factors contributing in the occurrence of an event. As, in our approach, factors are not necessarily rational decision-makers, we will propose a close model gathered in the notion of *incidence function* instead of that of characteristic function of cooperative game, and the *relevance measures* for factors instead of values for players. The events that can be analyzed with the proposed methodology include the analysis of factors in traffic accidents, in mining accidents, quality control analysis, diseases, etc.

We assume that the information available is the data obtained from the different repetitions of the event in the period under analysis, and a set of pre–selected factors that may intervene in its occurrence. From these data an *incidence function* a is defined. This incidence function associates to each subset of factors the number of times that these (and only these) factors were present in the occurrence of the event. Thus, the incidence function a encapsulates all the collected information on the event by assigning to each subset of factors the number of occurrences of the event in which precisely all these factors intervened.

It should be clarified that in our approach we are making three basic assumptions. First of all, we assume that no other information on the event, apart of the incidence function a, is available. Secondly, we suppose that the pre–selected factors are either clearly known or are assumed to be those identified as such by experts on each given event under study. Finally, the factors are assumed to be mutually independent, although in some cases this could not be totally true.

Once the incidence function a is obtained, the following step is to define an adequate measure to evaluate the relevance of each factor in a. Given an incidence function a, a *relevance measure* assigns a real number to each factor, stressing its level of importance in a. Among the relevance

measures we highlight one of them, the *fair measure*, for which we give a sound axiomatic foundation.

The relevance measures we propose for quantifying the importance of each factor can be used for different purposes such as: a budget distribution for improving the future occurrences of the event, a test for checking if previous policies were effective, or a preliminary study for performing subsequent complementary studies when some exogenous information not encapsulated in the incidence function a is available.

In the next subsection we introduce a simple but motivating example for a standard economical/university event: the distribution of a budget among researchers as a function of their past but immediate publication productivity. This is the only example we deal with in which factors are people. The section finishes by formally introducing the notion of incidence function.

The rest of the paper is organized as follows. In Section 2, the notion of relevance measure is introduced and a theoretical foundation is incorporated by means of some desirable properties that reinforce or weaken the measures presented. By using this methodology, one of them, the fair measure, is highlighted. Section 3 is focused on the fair relevance measure where two axiomatic characterizations are provided and a link with a well–known solution in cooperative games, the Shapley value, is provided. Section 4 treats two case studies by using real data involving accidents in road traffic and mining, respectively. Some final considerations conclude the paper.

1.1 The publication incidence function

To better motivate the utility of the tools, in this subsection we expose the example of the event of evaluating the scientific production in a university department (other examples can be found in [4]).

Periodically, a small university department shares a fraction of its resources among the members that have done scientific research in a fixed previous period of time according to their paper authorship in specialized journals. More precisely: if N is the set of researchers of the department then the total publication number $a(S)$ is assigned to each subset $S \subseteq N$ quantifying the papers coauthored by the members of S and published within the given period (e.g., $a(S)$ is the number of papers in journals appearing in the Journal Citation Reports of a given year). Once the function a is determined, we want to look for reasonable ways to assess funds to the different members of the department. For the sake of simplicity, we assume that the department has only four members, three senior researchers r, s and t, and a young researcher y, so that $N = \{r, s, t, y\}$. Assume, further, that the possible policies of the department are represented in the following scenarios, and assume that coauthorship with outsiders of the department is not taken into account:

1. Each published paper is rewarded equally and the spoil is equally divided among its authors.

2. Coauthorship within members of the department is stimulated and the spoil per paper is divided among authors equally.

3. Publication is encouraged for the young researcher, no matter the number of coauthors for her publications, and coauthorship with the young researcher is stimulated for the senior researchers.

Assume that the incidence function is:

$$\begin{aligned} a(r) = 10, \quad a(s) = 6, \quad a(t) = 2, \\ a(rs) = 0, \quad a(rt) = 2, \quad a(st) = 2, \quad a(sy) = 1, \quad a(ty) = 3, \\ a(sty) = 3, \\ a(\text{others}) = 0. \end{aligned}$$

From these data it may be observed that the three senior researchers are authors or coauthors in 12 papers each, while the young researcher is coauthor in 7 papers. The number of papers published by the department in the analyzed period is 29. Researcher r is by far the one who more contributed since he is the sole author of most of his publications. Researcher t is considerably less active in publishing papers alone or with department outsiders, but she is the most cooperative senior researcher over the other two. Researcher s plays an intermediate role.

- Scenario 1

We agree to score each published paper by 1, and equally dividing the score among the authors of the paper. Thus, the measure we are seeking is nothing else that:

$$\mathcal{F}_i(a) = \sum_{S \subseteq N : i \in S} \frac{a(S)}{|S|}, \quad i \in N$$

which in our particular case gives:

$$\mathcal{F}(a) = (11, 8.5, 6.5, 3)$$

and therefore the budget would be divided proportionally to these weights.

- Scenario 2

We may also consider a modified scheme for evaluating coauthored papers that provides incentives to collaboration among researchers. The score received by the authors of a paper is still equally divided among them, but now the score assigned to joint papers linearly increases with the number of coauthors. Indeed, this is a way to stimulate co-authorship within members of the group. Assume that an article with a single author is scored by 1 and that an extra score of ϵ is added to the papers for each additional author after the first one, i.e., the weights assigned to papers with 1, 2, 3 and 4 authors are $1, 1+\epsilon, 1+2\epsilon$ and $1+3\epsilon$, respectively. Then the measure is:

$$\mathcal{S}_i(a) = \sum_{S \subseteq N : i \in S} \frac{a(S)(1 + \epsilon(|S| - 1))}{|S|}, \quad i \in N$$

Let us consider in our example $\epsilon = 0.5$, so that the weights are 1, 1.5, 2, and 2.5 for papers of 1, 2, 3, and 4 authors respectively. The measure gives:

$$\mathcal{S}(a) = (11.5, 10.25, 9.25, 5).$$

The values obtained indicate the proportional scale to divide the budget among individuals according to the second criterion.

Consider now $\epsilon = 1$, in which case the weights assigned to papers with 1, 2, 3, or 4 authors are 1, 2, 3, and 4 respectively. The measure gives:

$$\tilde{\mathcal{S}}(a) = (12, 12, 12, 7)$$

which reflects the number of papers published by each author, independently of coauthorship.

We may consider also a modified scheme for evaluating coauthored papers that provides strong incentives for collaboration only among two researchers, i.e. the weight is 2 for papers with 2 authors and 1 in the other cases; the measure results to be:

$$\bar{\mathcal{S}}_i(a) = \sum_{S\subseteq N: i\in S, |S|\neq 2} \frac{a(S)}{|S|} + \sum_{S\subseteq N: i\in S, |S|=2} a(S), \quad i \in N$$

which gives

$$\bar{\mathcal{S}}(a) = (12, 10, 13, 5)$$

- Scenario 3

The criterion now does not treat symmetrically the seniors researchers when compared with the young researcher. All the publications by the young researcher are scored 1 for her, no matter the number of authors. Senior researchers are treated equally by the measure; for them the weights are $1/|S|$ whenever the publication has not the young researcher as coauthor, but the weights are $1/(|S|-1)$ if the young researcher is coauthor. That is, the senior researchers receive a better treatment by the measure if they cooperate with the young researcher. Then the measure is defined as follows:

$$\mathcal{P}_i(a) = \begin{cases} \sum_{S\subseteq N: i\in S, y\notin S} \frac{a(S)}{|S|} + \sum_{S\subseteq N: i, y\in S} \frac{a(S)}{|S|-1}, & \text{if } i \text{ is a senior researcher} \\ \sum_{S\subseteq N: i\in S} a(S), & \text{if } i \text{ is the young researcher} \end{cases}, \ i \in N$$

which gives

$$\mathcal{P}(a) = (11, 9.5, 8.5, 7).$$

Therefore the way to proportionally divide the budget according to the third criterion, which is by far the most beneficial for the young researcher.

To compare how the researchers are valuated by the previous measures we may refer to Table 1:

	r	s	t	y
$\mathcal{F}/29$	0.379	0.293	0.224	0.103
$\mathcal{S}/36$	0.319	0.285	0.257	0.139
$\tilde{\mathcal{S}}/43$	0.279	0.279	0.279	0.163
$\bar{\mathcal{S}}/40$	0.300	0.250	0.325	0.125
$\mathcal{P}/36$	0.306	0.264	0.236	0.194

Table 1: Normalized measures for the researchers

It is easy to see from Table 1 that $\mathcal{F}$ penalizes the young researcher y, who is favored by $\mathcal{P}$. On the other hand, $\mathcal{F}$ is the best option for r and s, while t would prefer $\bar{\mathcal{S}}$. Finally we may remark that s could be quite indifferent among the five measures.

1.2 Theoretical methodology

Let $\mathcal{P}$ be an event and $N = \{1, 2, \ldots, n\}$ be a selected set of significant independent factors[1] intervening in its occurrences. In this context, an incidence function on N is a function $a : 2^N \to$

[1]Factors should be interpreted in a broad sense, as for example components in a reliability system

$\mathbb{R}_{\geq}$ such that $a(\emptyset) = 0$.[2] Once the period of analysis of $\mathcal{P}$ is established, the *incidence function* a assigns to any subset S of N ($S \neq \emptyset$) the number of occurrences of $\mathcal{P}$ in which all the factors in S intervened, but none of the factors in $N \setminus S$. The data collected in each period of analysis generate the corresponding incidence function.

Note that, although usually the values $a(S)$ are represented by integer non-negative numbers, we also allow real non-negative values for the incidence functions. Due to this observation an incidence function is a cooperative game with non-negative images for the set of factors S, for which the set N are independent factors instead of decision-makers. Contrarily to what occurs in cooperative game theory, the function a by its nature fails to fulfill some properties promoting cooperation among decision-makers as monotonicity, convexity or superadditivity.[3]

For every finite set of factors N, we denote by $\mathcal{A}^N$ the class of all incidence functions on N. In the set $\mathcal{A}^N$, we can define two natural operations, the *sum* and the *product for a non-negative real number*, which give new incidence functions:

- If a_1, $a_2 \in \mathcal{A}^N$:
  $(a_1 + a_2)(S) = a_1(S) + a_2(S)$ for every set of factors $S \subseteq N$.
- If $a \in \mathcal{A}^N$ and $k \in \mathbb{R}_{\geq}$:
  $(ka)(S) = k \cdot a(S)$ for every set of factors $S \subseteq N$.

The set $\mathcal{A}^N$ with these operations has the structure of a cone in $\mathbb{R}^{2^N-1}$ with the null incidence function η defined by $\eta(S) = 0$ for all set of factors $S \subseteq N$ as proper zero element in $\mathcal{A}^N$. The cone $\mathcal{A}^N$ is generated by the vectors associated to the incidence functions $a_S \in \mathcal{A}^N$ given by

$$a_S(R) = \begin{cases} 0 & \text{if } R \neq S \\ & \text{, for all } \ S \subseteq N, S \neq \emptyset \\ 1 & \text{if } R = S \end{cases} \tag{1}$$

Consequently, using the above introduced operations, we can now write any $a \in \mathcal{A}^N$ as

$$a = \sum_{S \subseteq N:\ S \neq \emptyset} a(S) \cdot a_S$$

Finally, given an incidence function $a \in \mathcal{A}^N$, we will sometimes consider the total number of occurrences of $\mathcal{P}$ in that period of time:

$$T(a) = \sum_{S \subseteq N} a(S)$$

2 Relevance measures

This section is divided into two subsections. In the first one we define the concept of relevance measure and show some examples. In particular, we introduce a very intuitive relevance measure: the fair measure. In the second subsection we define some properties that can be verified for a relevance measure and establish some relations among them.

[2] We exclude that there exist situations that involve no factor.

[3] We will consider these properties later on.

2.1 Definition and examples

A relevance measure provides an evaluation on how significant a factor is for an event described *only* by an incidence function.

Definition 2.1 ***Relevance measure***
A relevance measure *is a function* $f : \mathcal{A}^N \to \mathbb{R}^N_{\geq}$ *that assigns to every incidence function,* a*, the vector* $(f_1(a), f_2(a), \ldots, f_n(a))$ *where the non-negative real number* $f_i(a), i \in N$ *is interpreted as the importance of factor* i *in the event associated to the incidence function* a*.*

Different relevance measures can be defined on an incidence function a. We want to point out that the examples of relevance measures we are going to show are quite natural in the sense that someone not necessarily trained in the design of measures might consider them intuitively as natural measures for evaluating the significance of factors in events described by incidence functions.
In what follows let $a \in \mathcal{A}^N$ be any incidence function and $i \in N$ any factor.

Example 2.2 The egalitarian measure $\mathfrak{e}$

$$\mathfrak{e}_i(a) = T(a)/n$$

The egalitarian measure assigns the same value to all factors, independently of the frequencies with which they appear. It is a solidarity measure.

Example 2.3 The basic measure $\mathfrak{b}$

$$\mathfrak{b}_i(a) = \sum_{S \subseteq N \,:\, i \in S} a(S)$$

The basic measure is the second one proposed in scenario 2 in Subsection 1.1. This measure seems very natural when we suppose that any factor is able to generate the outcome even independently from the others.

Example 2.4 The fair measure $\mathfrak{F}$

$$\mathfrak{F}_i(a) = \sum_{S \subseteq N:\, i \in S} \frac{a(S)}{|S|}$$

The fair measure is the one proposed in scenario 1 in Subsection 1.1. This is the natural measure to be chosen if all factors in each set are supposed to have the same a priori weight and each occurrence of the event is treated equally.

Example 2.5 The weighted measures $\mathfrak{b}^c$

$$\mathfrak{b}^c_i(a) = \sum_{S \subseteq N:\, i \in S} a(S) c(i, S)$$

where $c : N \times 2^N \to \mathbb{R}$ is a function which allows to weight subsets in a different way for any $i \in N$.

The basic and the fair measures are particular cases of these measures when $c(i,S) = 1$ and $c(i,S) = \dfrac{1}{|S|}$, respectively, for any $i \in N$. As we shall see, all the measures considered in Subsection 1.1 are of this kind.

Example 2.6 The selective measures $\mathfrak{s}^\alpha$

$$\mathfrak{s}_i^\alpha(a) = \sum_{S \subseteq N\,:\, i=\alpha(S)} a(S)$$

where α is a selection function, $\alpha : 2^N \to N$, with $\alpha(S) \in S$ for all $S \neq \emptyset$.

This measure seems very natural when for each set of factors S we are able to assign the whole importance to just one factor, namely $\alpha(S)$. In a sense the other factors in S, if any, are depending on $\alpha(S)$.

Notice that a selective measure can also be viewed as a weighted measure in which the weights are defined by

$$c(i,S) = \begin{cases} 1 & \text{if } i = \alpha(S) \\ 0 & \text{otherwise} \end{cases}$$

Example 2.7 The proportional measure $\mathfrak{p}$

$$\mathfrak{p}_i(a) = \frac{T(a)}{\sum\limits_{j=1}^{n} a(\{j\})} \cdot a(\{i\})$$

The proportional measure is well–defined if in at least one occurrence of the event only a single factor appeared. This measure seems very natural when we are not sure that when a occurrence involves more than one factor all the factors are really effective; we may think to a road accident that involves a driver with serious damages on a car in bad condition, but we are not able to say if these negative elements where already present before the accident, so that they are the effective factors, or one of the two is simply a consequence of the accident. In this case we may take into account just those occurrences that depend on a single factor and measure the relevance proportionally to it.

These examples are just different ways to design a relevance measure. The adequate relevance measure to use in each event $\mathcal{P}$ under analysis must be designed according to the objectives of this analysis. For example, for the scenario 3 in Subsection 1.1, an "ad hoc" measure has been defined to differentiate between young and senior researchers. As a general comment, we may say that the egalitarian measure, being a solidarity measure, flattens the differences among the factors, so it may be used whenever the reliability of the data is extremely low; the basic measure is suitable when the concurrency of the factors is negligible, i.e. any factor may generate the outcome even alone; the fair measure is useful when all the set of factors and all the factors in each set may be considered of equivalent weight; the selective measure emphasizes the importance of just one factor for each set; the proportional measure is the most suitable when the concurrency of more than one factor is viewed as a "noise", so it is preferable to ignore the outcomes involving more than one factor. The practice and the knowledge of a manager may allow defining more suitable measures.

Three aspects concerning relevant measures should be highlighted here:

First, relevance measures cover a wide spectrum of tools to be used for measuring the importance of factors. Depending on the policy that the evaluators (usually experts on the analyzed

field) seek, they will choose one measure or another. Measures can range from individualistic criteria favoring the strongest factors to solidarity criteria (e.g., observe that the different measures adopted in the example in subsection 1.1, lead to different policies). This wide range of possibilities is best reflected at the end of section 3.

Second, once the measure is chosen by experts we want to emphasize that the only information needed to compute them is the incidence function. This is an advantage with respect to some alternative methods for measuring the importance of factors that use some extra data.

Third, these measures are easy to be computed. From the computational point of view, the only limitation is the number of factors: whenever the incidence function is treatable most of the chosen measures can be computed. In other words, the complexity of the problem uniquely depends on the input data if the measure is defined by elementary operations, as it is the case of all the particular measures considered in this paper.

2.2 Some properties for relevance measures

In this subsection we introduce two properties for factors and five properties for relevance measures that will be used later.

Definition 2.8 *Let $a \in \mathcal{A}^N$.*

- *A factor i is* null *in a if $a(S) = 0$ for all $S \subseteq N$ with $i \in S$.*
- *Two different factors i and j have* equivalent incidence *in a if $a(S \cup \{i\}) = a(S \cup \{j\})$ for all $S \subseteq N \setminus \{i, j\}$.*

Definition 2.9 *A relevance measures f satisfies the property of:*

- Totality: *if $\sum_{i \in N} f_i(a) = T(a)$ for all $a \in \mathcal{A}^N$.*
- Zero on nulls: *if $f_i(a) = 0$ for any factor i null in $a \in \mathcal{A}^N$.*
- Equal treatment: *if $f_i(a) = f_j(a)$ for all pairs of factors i and j with equivalent incidence in $a \in \mathcal{A}^N$.*
- Linearity: *if $f_i(\alpha a + \beta b) = \alpha f_i(a) + \beta f_i(b)$ for all $\alpha, \beta \in \mathbb{R}_{\geq}, a, b \in \mathcal{A}^N$ and for all $i \in N$.*
- Monotonicity: *if $a(S) \geq b(S)$ for all $S \subseteq N, S \ni i$ implies $f_i(a) \geq f_i(b)$ for all $a, b \in \mathcal{A}^N$.*

Linearity allows for weighted combinations of different incidence functions. For instance, referring to the example in Subsection 1.1 we can make use of the publications in different periods, with weights that allow to differently account the various periods.

Monotonicity tells that if a factor i has larger incidence in function a than in function b, then the relevance for factor i in function b should be at most the same as in function a.

In the following table we resume which of these properties are verified in the examples given in the former subsection, leaving the proofs as an exercise for the reader.

	Totality	Zero on nulls	Equal treatment	Linearity	Monotonicity
Egalitarian	Yes	No	Yes	Yes	No
Basic	No	Yes	Yes	Yes	Yes

Fair	Yes	Yes	Yes	Yes	Yes
Weighted	No*	Yes	No*	Yes	Yes
Selective	Yes	Yes	No	Yes	Yes
Proportional	Yes	Yes	Yes	No	No

Table 2: Some properties of relevance measures

In this table No* means that the property is verified or not depending on the considered weights. Notice that there is only one of the former measures which verifies all of these properties: the fair relevance measure defined in Example 2.4. Similarly to our approach, properties and characterization of risk measures by means of axioms is a common technique for risk measures in insurance markets [33].

The properties we consider are not independent. In the following proposition we prove some links among them.

Proposition 2.10 *Let f be a relevance measure.*

- *If f verifies Totality, Monotonicity and Equal treatment properties then it also verifies Zero on nulls property.*
- *If f verifies Totality, Monotonicity and Equal treatment properties then it also verifies Linearity property.*

3 The fair relevance measure

In this section we go deeper in the analysis of the fair relevance measure. The special interest for this measure arises as it satisfies all the properties introduced in Subsection 2.2, and it seems also very intuitive.

Apart of the interest of its properties, the fair measure is balanced and not biased in the sense that it fully reflects the concurrence of each factor in the event without favoring the most relevant factors and non-being solidary with the least relevant factors.

Recall that the fair relevance measure was defined in Example 2.4 by:

$$\mathfrak{F}_i(a) = \sum_{S \subseteq N:\, i \in S} \frac{a(S)}{|S|}, \tag{2}$$

for each factor $i \in N$ and for any $a \in \mathcal{A}^N$. The value $\mathfrak{F}_i(a)$ is the result of adding up, for each set S containing the factor i, the contribution of this factor i in $a(S)$, assuming that all factors in S have the same numerical assignment.

As was shown in Table 2, the fair relevance measure satisfies Totality, Zero on nulls, Equal treatment, Linearity and Monotonicity properties. We see now that some of these properties characterize this measure.

Proposition 3.1 *There exists only one relevance measure that satisfies: Totality, Zero on nulls, Equal treatment and Linearity properties. This measure is precisely the fair relevance measure.*

The four properties considered in Proposition 3.1 constitute a *minimal* characterization of the fair relevance measure. To prove this minimality we show, for each one of these four properties, an example, taken from Table 2, of relevance measure that satisfies the three other properties, but *not* the property in question.

Example 3.2

- The basic measure $\mathfrak{b}$, defined in Example 2.3, satisfies Zero on nulls, Equal treatment and Linearity but it does not verify Totality.
- The egalitarian measure $\mathfrak{e}$, defined in Example 2.2, satisfies Totality, Equal treatment and Linearity, but it does not satisfy Zero on nulls.
- The selective measure $\mathfrak{s}^\alpha$, defined in Example 2.6, satisfies Totality, Zero on nulls and Linearity, but it does not satisfy Equal treatment.
- The proportional measure $\mathfrak{p}$, defined in Example 2.7, satisfies Totality, Zero on nulls and Equal treatment, but it does not satisfy Linearity.

We can provide another more difficult characterization of the fair relevance measure.

Theorem 3.3 *There exists only one relevance measure that satisfies: Totality, Equal treatment and Monotonicity properties. This measure is precisely the fair relevance measure.*

This previous theorem is an immediate consequence of Propositions 2.10 and 3.1.

The three properties considered in theorem 3.3 constitute another *minimal* characterization of the fair relevance measure. To prove this minimality we show for each of these properties an example that satisfies the two other properties, but *not* the property in question.

Example 3.4

- The basic measure $\mathfrak{b}$ defined in Example 2.3 satisfies Equal treatment and Monotonicity, but it does not verify Totality.
- The selective measure $\mathfrak{s}^\alpha$ defined in Example 2.6 satisfies Totality and Monotonicity but it does not satisfy Equal treatment.
- The proportional measure $\mathfrak{p}$, defined in 2.7, satisfies Totality and Equal treatment but it does not satisfy Monotonicity.

3.1 The fair measure seen under the viewpoint of cooperative game theory

Formally a *cooperative game* is a pair (N, v) where N is a finite set of players and $v : 2^N \to \mathbb{R}$ is the characteristic function, that assigns a real number $v(S)$ to each coalition $S \subseteq N$, with $v(\emptyset) = 0$. Thus an incidence function, as defined in subsection 1.2, can be viewed as a cooperative game in which N, the set of players, admits a broader context in which players can be replaced by factors, being possibly non-decision-makers, i.e., the factors act as 'black boxes' with no decision–making ability. This difference becomes crucial.

Cooperative games have properties that encourage cooperation among players and from these properties it results the problem of how to share the total amount raised among the players. Three of these, non-independent properties are:

- *Monotonicity* if $v(S) \leq v(T)$ for all $S \subset T$.
- *Convexity* if $v(S \cup T) + v(S \cap T) \geq v(S) + v(T)$, for all $S, T \subseteq N$.
- *Superadditivity* if $v(S \cup T) \geq v(S) + v(T)$ for all $S, T \subset N$ such that $S \cap T = \emptyset$.

It is clear that if a game is convex then it is superadditive and monotonic. Thus, in a convex cooperative game the rational decision-makers will form the grand coalition N and after they will discuss how to divide the spoil $v(N)$ among them. The most well-recognized one-point solution that gives the answer to the second issue is the Shapley value.

On the other hand, it is clear that by its nature the incidence functions we are considering are neither monotonic, nor convex, nor superadditive. To remedy this we may consider the *cumulative incidence function* v_a regarded as a cooperative game, instead of a, where $v_a(S)$ counts the total number of occurrences of the event in which any combination of (not necessarily all) factors in S concurred, but none of the factors in $N \setminus S$, i.e.,

$$v_a(S) = \sum_{R \subseteq S} a(R) \qquad \text{for all } S \subseteq N \tag{3}$$

The correspondence between v_a and a is one–to–one since the determinant associated to the equations system in (3) is 1, so we may derive v_a from a and conversely.[4]

The next result proves that the cumulative incidence function v_a is convex and therefore monotonic and supperadditive. Thus it has all the ingredients for taking the total amount $v_a(N)$ and dividing it among factors according to the Shapley value.

Proposition 3.5 *Given any incidence function $a \in \mathcal{A}^N$ the function v_a defined in* (3) *is a convex cooperative game.*

Shapley [32] proved that there exists a unique function (value), on the set of cooperative games having players' set N, which satisfies: Efficiency, Null player, Symmetry and Linearity (the *Shapley value*, Φ) which is given by

$$\Phi_i(v) = \sum_{S \subseteq N \setminus \{i\}} \frac{s!(n-s-1)!}{n!} (v(S \cup \{i\}) - v(S)) \tag{4}$$

where $s = |S|$ and a value ψ for cooperative games is a function which assigns a real number $\psi_i(v)$ to any game v and any player $i \in N$. The four properties are the following.

- Efficiency: $\sum_{i \in N} \psi_i(v) = v(N)$ for any game v.
- Null player: $\psi_i(v) = 0$ if $v(S \cup \{i\}) = v(S)$ for all $S \subseteq N \setminus \{i\}$.
- Linearity: $\psi_i(\alpha v + \beta w) = \alpha \psi_i(v) + \beta \psi_i(w)$.
- Anonimity: $\psi_{\pi(i)}(\pi(v)) = \psi_i(v)$.

where we are assuming that i is any player in N, v, w any games on N and α, β any real numbers), and $\pi : N \to N$ a permutation that preserves the game, i.e. $\pi(v(S)) = v(\pi(S))$ for any $S \subseteq N$.

The following significant relationship holds.

[4] a is derived from v_a by $a(S) = \sum_{R \subseteq S} (-1)^{s-r} v_a(R)$ in which $s = |S|$ and $r = |R|$.

Proposition 3.6 *Let $\mathfrak{F}$ be the fair relevance measure* (2) *and Φ be the Shapley value* (4) *. Then,*

$$\mathfrak{F}(a) = \Phi(v_a)$$

where v_a is the cumulative function defined from the incidence function a in (3)*.*

This result tells us that the fair measure applied to an incidence function is nothing else than the Shapley value applied to the associated cumulative incidence function v_a.

Considering the cumulative incidence function v_a is, in our context, less natural than using the incidence function a. Moreover, we remark that it is preferable computing $\mathfrak{F}$ to a than Φ to v_a since our data is a and therefore we avoid the step of getting v_a from a. Additionally, the formula to compute $\mathfrak{F}$ is simpler than the one for Φ which is computationally intractable when the number of factors is considerably high.

The relationship we have just established between the Fair relevance measure of an incidence function and the Shapley value for the corresponding cumulative incidence function can be extended to other relevance measures, in the sense that they correspond to other well–known measures in cooperative game theory. For instance, it can easily be proved that the weighted relevance measure (see definition 2.5) with $c(i,S) = \dfrac{1}{2^{|S|-1}}$ for all $i \in N$ and $S \subseteq N$ corresponds to the Banzhaf value (defined by Owen [30]) of the corresponding cumulative incidence function. Also, the first two measures proposed in scenario 2, in the introductory example of subsection 1.1, correspond to certain semivalues of the cumulative incidence function, and even the measure proposed in scenario 3 corresponds to a probabilistic value of the cumulative incidence function. However, this kind of relationship does not always exist: the third relevance measure considered in scenario 2 does not correspond to any probabilistic value for the cumulative incidence function. In summary, many relevance measures considered in section 2 have their analogues in game theory by considering the cumulative incidence function, but some other do not have this correspondence.

Beyond game theory, our approach can be related to measures used in different fields such as customer satisfaction ([7]), systems reliability ([19]) or medical research. For instance, two of the measures considered in attributable risk analysis, the approach known in statistics for medical research ([18], [17]), can also be seen from our point of view: the attributable risk (AR) and the relative risk (RR). The AR measure for a risk factor i (AR_i) is the probability of a member of a population having the disease ($P(D)$) reduced by the probability of having the disease without the risk factor ($P(D|i^-)$), relative to ($P(D)$):

$$AR_i = \frac{P(D) - P(D|i^-)}{P(D)} = 1 - \frac{P(D|i^-)}{P(D)}$$

and the RR measure for a risk factor i is defined as the number

$$RR_i = \frac{P(D|i^+)}{P(D|i^-)}$$

where $P(D|i^+)$ is the probability of having the disease presenting the risk factor.

Both, AR and RR, are taken as additional measures for evaluating the importance of risk factors in an event and are usually estimated, in empirical research, as proportions from the sample data (see e.g. [7]). In our context, given the sample data, if we consider the incidence function a which assigns to each subset S of the set N of selected risk factors the number of times that all factors in the subset were present in a person having the disease, it would be

natural to consider the two respective analogues for AR and RR, which are particular examples of relevance measures for the risk factors:

$$AR'_i = 1 - \frac{\sum\limits_{S\subseteq N, i\notin S} a(S)}{\sum\limits_{S\subseteq N} a(S)}$$

$$RR'_i = \frac{\sum\limits_{S\subseteq N, i\in S} a(S)}{\sum\limits_{S\subseteq N, i\notin S} a(S)}$$

4 Two case studies involving risk factors

We introduce two cases of study involving sinistrality. In both situations the available information taken from official reports just let us to build the incidence function a. As the fair relevance measure is the only one that treats each occurrence of the event and also each factor equally, it seems to be the relevance measure far better placed to be used. Although the following examples just consider some few factors it should be noted that the number of factors could be very high. For instance, in medicine it is an important problem to establish a ranking of importance among genes causing a given cancer. Risk factors are genes provoking the illness and can be several thousands, see e.g. [24].

4.1 Determining the importance of factors in traffic accidents

Road traffic deaths constitute a major public problem and are predicted to increase if road safety is not addressed adequately. The appalling human misery and the serious economic loss caused by road accidents demand the attention of the society and call for higher levels of safety. Traffic safety is usually regarded in terms of traffic "unsafety", i.e., as the number of fatalities or injuries resulting from traffic accidents.

In the case study we present in this subsection we center our attention in five factors intervening in mortal road traffic accidents (RTA) which, at the same time agglutinate several subcases: (1) breaking traffic laws (ignoring traffic signals, safety belt not fastened, overtaking when not allowed...), (2) driver's errors (overtaking without visibility, not being aware of external circumstances...), (3) inappropriate state of the driver (alcohol or drugs effects, illness, distractions,...), (4) inadequate speed, (5) others (faulty state of the vehicle, faulty state of the road, fortuitous unforeseeable external causes,...). The five factors considered are the most relevant ones mentioned in the annual statistic data of mortal RTA in Catalonia in 2009, where concurrent factors were reported.

If $N = \{1, 2, 3, 4, 5\}$ is the fixed set of factors considered, then, for each mortal RTA (in the context under analysis) we have a partition of N into two subsets S and $N \setminus S$, understanding that the factors in S were all present in the RTA and none of the elements in $N \setminus S$ intervened. For instance, if $S = \{1, 3\}$ is assigned to a particular RTA it means that breaking a traffic law and inappropriate state of the driver concurred in it, while the other three factors were absent. Then, the incidence function corresponding to these data $a : 2^N \to \mathbb{R}$ is such that $a(S)$ (for $S \neq \emptyset$) gives the number of RTA occurred with the sole presence of (all factors of) S and without the intervention of any factor in $N \setminus S$. We assume $a(\emptyset) = 0$.[5]

[5]We do not count the RTA with no reported factor.

The incidence function a given by the collected data is the following:

$a(\{1\}) = 1323$, $a(\{2\}) = 1432$, $a(\{3\}) = 1517$, $a(\{4\}) = 496$, $a(\{5\}) = 172$,
$a(\{1,2\}) = 229$, $a(\{1,3\}) = 176$, $a(\{1,4\}) = 63$, $a(\{1,5\}) = 68$,
$a(\{2,3\}) = 0$, $a(\{2,4\}) = 116$, $a(\{2,5\}) = 60$, $a(\{3,4\}) = 43$,
$a(\{3,5\}) = 4$, $a(\{4,5\}) = 74$, $a(\{1,2,3\}) = 44$, $a(\{1,2,4\}) = 0$,
$a(\{1,2,5\}) = 0$, $a(\{1,3,4\}) = 2$, $a(\{1,3,5\}) = 0$, $a(\{1,4,5\}) = 0$,
$a(\{2,3,4\}) = 0$, $a(\{2,3,5\}) = 0$, $a(\{2,4,5\}) = 0$, $a(\{3,4,5\}) = 0$,
$a(\{1,2,3,4\}) = 0$, $a(\{1,2,3,5\}) = 0$, $a(\{1,2,4,5\}) = 0$, $a(\{1,3,4,5\}) = 0$,
$a(\{2,3,4,5\}) = 0$, $a(N) = 0$.

The total number of accidents $T(a) = \sum_{S \subseteq N} a(S) = 5819$.

The computation of the proportional measure (see Definition 2.7) gives

$$\mathfrak{p}(a) = (1558.4, 1686.8, 1786.9, 584.3, 202.6),$$

while the basic measure (see Definition 2.3) is

$$\mathfrak{b}(a) = (1905, 1881, 1786, 794, 378),$$

and the fair measure (see Definition 2.4) gives

$$\mathfrak{F}(a) = (1606.3, 1649.2, 1643.8, 644.7, 275).$$

From these results we can derive several significant considerations.

- The ranking of relevance given by the proportional measure $\mathfrak{p}$ orders the five factors as follows: $3 > 2 > 1 > 4 > 5$; the ranking of relevance given by the basic measure $\mathfrak{b}$ is: $1 > 2 > 3 > 4 > 5$; and the ranking of relevance given by the fair measure $\mathfrak{F}$ is: $2 > 3 > 1 > 4 > 5$. Hence, the three rankings are different. Factor 2 turns up to be more important than factor 3 for $\mathfrak{b}$ and for $\mathfrak{F}$. Note that the proportional measure wrongly tells that factor 3 is more relevant than factor 2. This is due to the fact that the proportional measure disregards the mortal RTA with concurrence of several factors. The basic measure ranks factor 1 as the most important because it is the most frequently reported factor concurrent with others, although it is not the most frequent as a single causer of an RTA.

- Assume that a budget has to be invested in actions addressed to prevent sinistrality. How should this total amount be distributed in proportion to the factors that provoke sinistrality? In our opinion, supported by the results given in previous sections, we would give support to the distribution given by the fair relevance measure $\mathfrak{F}$. Unfortunately, the proportional measures, which omits concurrence of several factors, is occasionally wrongly taken.

- Assume, again, that for some reason it is decided that the budget is going to be invested in actions addressed to prevent sinistrality related to just two factors. Which factors should be chosen? In this situation, $\mathfrak{F}$ would propose to share the total budget between factors 2 and 3 in an almost equal proportion.

If accidents police reports just provide contributing factors in each accident in binary form (either yes or no) as is the case, little more can be done to calculate the incidence function and

compute for it a reasonable relevance measure for it. Of course, if we had more information this should be incorporated in our model and treat the problem in a different way.

Independence of factors is another significant issue, however there exist some literature of several techniques for the Shapley value (see, e.g. [11] and [26]) which can be adapted to our context, but this is out of our scope in this paper.

4.2 Determining the importance of factors in mining accidents

Decisions in mining safety management are typically about actions (interventions, projects, regulations, standards, programs, etc.) that involve the expenditure of money to reduce the toll of accidents. The central feature of such decisions is that they are made by few on behalf of many. To have legitimacy one must be able to claim that what the few decide is what the many give their consent to.

The data in this section are obtained from the "Ministerio de Industria, Energía y Minería" and correspond to all mining accidents occurred in Spain between 1982 and 2006. The codification used to describe the factors (in this context they are called events) intervening in a particular accident is the one recommended by experts in mining safety [8], and is described for this particular case in [31].

For each accident, a sequence of up to three temporally ordered events, that immediately preceded it, are reported. They are called *precursor events*. All precursor events are classified by two complementary characteristics: the type of event and the temporal order in which they occurred. This last characteristic is called the *precursor level of the event*, and can be 1, 2 or 3, meaning that all events of level 1 took place before of the events of level 2, and these ones happened before the ones of level 3. In this way we obtain the following classification:

1. **Environmental events**: events referring to the location of the accident (e.g., low lighting, wet floor, or cramped conditions). Denoted by V1, V2, or V3, depending on their precursor level.

2. **Equipment events**: events resulting from breakage or malfunction of machinery or tools. Denoted by E1, E2, or E3.

3. **Medical events**: events resulting from the person's current state of physical well-being (e.g., heart attack or diabetic or epileptic episode). Denoted by M1, M2, or M3.

4. **Behavioral events**: events resulting directly from human involvement (e.g. learning too far into the path of machinery, touching an electrically charged object). Denoted by B1, B2, or B3.

Let N be the finite set of precursor events. The cardinality of N is 12 and its elements are:

$$\{V1, V2, V3, E1, E2, E3, M1, M2, M3, B1, B2, B3\}$$

Let $a : 2^N \to \mathbb{R}$ be the incidence function such that $a(S)$ gives the number of mining accidents that have occurred in Spain during the analysis period with the sole presence of all the events in S, i.e., without the intervention of any element of $N \setminus S$. If T is the total number of accidents, we have

$$\sum_{i \in S} a(S) = T.$$

Some subsets of N technically have null incidence. Indeed, $a(S) = 0$ whenever the cardinality of S is greater than 3 (recall that at most three precursor events are reported for each accident), or whenever S contains an event of the same type in more than one precursor level, e.g. the subset $S = \{V1, E2, V3\}$ cannot have positive incidence. If only one event occurred in an accident, it must obviously be the "first" one so that subsets like $S = \{V2\}$ or $S = \{V3\}$ are not possible. Similarly, if two events occurred in an accident they must be consecutively the "first" and the "second" but never the third, e.g. subset $S = \{V1, B3\}$ is neither possible. Finally, we also assume the technical requirement that $a(\emptyset) = 0$, i.e. if no event is known for an accident, something that occurs two times in our data, we do not count this accident. The total number of accidents taken into account is $T = 242$.

As the cardinality of the set of events N is 12 there are in principle $2^{12} = 4096$ subsets to take into consideration, but if we discard the ones with null incidence explained above, we only need to consider 40 subsets with potential non-null accident incidence. We do not reproduce here the values of the incidence function a on these 40 subsets, but using the fair relevance measure we obtain the following result (relevance of each event in N):

V1	V2	V3	E1	E2	E3	M1	M2	M3	B1	B2	B3	Total
93.83	4.33	0.67	22.33	4.67	1.00	67.83	42.00	4.00	1.00	0.00	0.33	242

Rescaling the data to one hundred we have the percentages of relevance of each event.

V1	V2	V3	E1	E2	E3	M1	M2	M3	B1	B2	B3	Percentage
38.77	1.79	0.28	9.23	1.93	0.41	28.03	17.40	1.65	0.41	0.00	0.14	100

Four events protrude above the other events:

$$V1 > M1 > M2 > E1 > others$$

In particular these four events represent more than the 93% of the total relevance. Thus, any effort to prevent future sinistrality should mainly be addressed to these four events.

If we do not make distinctions among the ordering of occurrence of events and only consider the four general types: V, E, M and B, understanding that V agglutinates $V1$, $V2$ and $V3$. Then

V	E	M	B	Total
98.83	28.00	113.83	1.33	242

Note that while clearly $V1 > M1$, the opposite $M > V$ holds. This is because $V1/V$ is close to 1, which means that event V rarely intervenes as a second or third temporarily event, and also because M is mainly distributed between $M1$ and $M2$.

As we have seen, experts on mining have grouped risks into these four categories although there exists a large sublist of factors within behavioral events for which our computation can be extended. This form of grouping factors is common in civil engineering, according to [5] more than 500 risk issues related to different stages in tunnel projects are identified. Nevertheless, some of them (six reported in [5]) represent the ones that were most frequently selected by experts as being major risks during construction.

5 Conclusion

In this paper we presented a game theoretic approach for analyzing the importance of factors in the occurrences of an event. We suppose that, for each subset of factors, the number of

occurrences of the event in which exactly these factors intervened is known. This available dataset is expressed by an incidence function. We provide adequate measures to evaluate the relevance of each factor in the occurrence of an event, that can be used for different policies in order to improve future repetitions of the event. We introduce some properties, analyze some relationship among them and mention similarities with measures in other contexts. Different relevance measures can be used depending on the problem under analysis and on the pursued goals. A part from the ones defined in Section 2, the experience and expertise of a manager may suggest further less intuitive but more appropriate measures, accounting also the situation at hand and the scope of the study.

We theoretically highlight a measure that treats equally all occurrences of the event and also all factors, the fair relevance measure. We relate it with a well-known value for cooperative games, the Shapley value, and show how this measure can be applied to situations involving risk factors in two types of accidents: road traffic and mining. The Shapley value for cooperative games is well established and accepted. If we had to select the greatest criticism that has received we would choose that it does not favor solidarity among the decision-makers. This is reflected in the first scenario of our initial example, in which the young researcher is treated as the other (senior) researchers. In our context we mainly have factors different from people thus, why should solidarity be an important issue? In our view the Shapley value is reinforced in the context considered in the paper.

Throughout the paper, we supposed that the factors are independent, but sometimes this hypothesis may result too strong. More precisely, it is possible that factors considered independent are connected in practice. Game theory offers a tool for facing these situations. We may refer to [26] where a situation of correlation among genes is analysed (see also [28] and [11]). In particular, they consider a game where the players are the genes that are supposed fully independent and their importance as markers of a disease is evaluated via the Shapley Value [32]. Following the approach in [29] a graph is added, in order to emphasize the existing relationships among the genes, and again their importance is evaluated via the Myerson Value. The difference between the Shapley Value and the Myerson Value assigned to the genes gives a measure of the centrality (see [16]) of the genes in the network and may be interpreted as a measure of their interdependence.

Possible future developments may look for exploiting the cooperative game introduced in Section 3 for analyzing possible dependencies among the factors. Another interesting direction is represented by those situations in which several factors are identified, so that the number of possible subsets increases and it may become difficult to manage the incidence function, so we can study the possibility of using approximated measures. A third situation that may deserve further investigation may show up when we consider many factors, and consequently the reliability of the data is questionable, as it is possible that not all the factors involved may be identified with enough certainty. In these cases it may be useful to use a subset of the available data, e.g. only those related to occurrences that are caused by at most two factors.

From a theoretical viewpoint it would be interesting to extend Proposition 3.6 of this paper and its following remarks by studying possible relationships between the class of weighted relevance measures (Example 2.5) and some well-known one-point solution concepts in cooperative game theory.

Acknowledgements

The ideas of the paper were discussed, and the paper itself was prepared mostly during some exchange visits of Vito Fragnelli and Josep Freixas. Both are grateful to the hosting departments for their warm hospitality. They acknowledge a grant from GNAMPA, CNR, supporting the visit in Italy of the second author. This research was partially supported by Grants SGR 2009–1029 of "Generalitat de Catalunya" and MTM2012-34426/FEDER of "Ministerio de Economía y Competitividad".

The authors wish to thank the referees for their comments and suggestions which contributed to improve the original version of this paper.

References

[1] J.M. Alonso-Meijide and C. Bowles. Generating functions for coalitional power indices: An application to the IMF. *Annals of Operations Research*, 137:21–44, 2005.

[2] T. Aven and R. Østebø. Two new component importance measures for a flow network system. *Reliability Engineering*, 14:75–80, 1986.

[3] P.J. Boland and E. El-Neweihi. Measures of component importance in reliability theory. *Computers Ops Res*, 4:455–463, 1995.

[4] F. Carreras and J. Freixas. Semivalue versatility and applications. *Annals of Operations Research*, 109:341–356, 2002.

[5] I. Chivatá Cárdenas, S.S.H. Al-jibouri, J.I.M. Halman, and F.A. von Tol. Capturing and integrating knowledge for managing risks in tunnel works. *Risk Analysis*, 33:92–108, 2013.

[6] M. Conklin and S. Lipovetsky. Marketing decision analysis by TURF and Shapley value. *Information Technology and Decision Making*, 4: 5–19, 2005.

[7] M. Conklin, K. Powaga and S. Lipovetsky. Customer satisfaction analysis: identification of key drivers. *European Journal of Operational Research*, 154: 819–827, 2004.

[8] A. Feyer and A.M. Williamson. A classification system for causes of occupational accidents for use in preventive strategies. *Scandinavian Journal of Work and Environmental Health*, 17:302–311, 1991.

[9] V. Fragnelli, I. García-Jurado, H. Norde, F. Patrone and S. Tijs. How to share railway infrastructure costs? In F. Patrone, I. García-Jurado, S. Tijs, editor, *Game Practice: Contributions from Applied Game Theory*, 91–101. Kluwer - Amsterdam (NL), 1999.

[10] V. Fragnelli and M.E. Marina. An axiomatic characterization of the Baker-Thompson rule. *Economics Letters*, 107: 85–87, 2010.

[11] V. Fragnelli and S. Moretti. A game theoretical approach to the classification problem in gene expression data analysis. *Computers and Mathematics with Applications*, 55:950–959, 2008.

[12] J. Freixas and M. Pons. Identifying optimal components in a reliability system. *IEEE Transactions on Reliability*, 57:163–170, 2008.

[13] G. Gambarelli and I. Stach. Power indices in politics: Some results and open problems. *Homo Oeconomicus*, 26:417–441, 2009.

[14] D.B. Gillies. *Some Theorems on n-person Games*. Ph.D. dissertation, Princeton, Princeton University Press, 1953.

[15] D.B. Gillies. Solutions to general non-zero-sum games in contributions to the theory of games. In Luce R.D. Tucker A.W., editor, *Annals of Mathematics Studies 40*, volume IV, 47–85. Princeton University Press, Princeton, USA, 1959.

[16] D. Gómez, E. González-Arangüena, C. Manuel, G. Owen, M. del Pozo and J. Tejada. Centrality and power in social networks: a game theoretic approach. *Mathematical Social Sciences*, 46:27–54, 2003.

[17] M.J. Kahn, W.M. O'Fallon and J. Sicks. Epidemiologic research: Principles of quantitative methods. *Lifetime Learning Publications*, Belmont, CA., 1982

[18] D.G. Kleinbaum, L.L. Kupper and H. Morgenstern. Generalized population attributable Risk Estimation. Technical report 54. *Department of health sciences research, Mayo Clinic, 2000.*

[19] W. Kuo and X. Zhuo. *Importance measures in reliability, risk, and optimization.* Wiley, 2012.

[20] S. Lipovetsky. SURF - structural unduplicated reach and frequency: latent class TURF and Shapley value analyses. *Information Technology and Decision Making*, 7: 203–216, 2008.

[21] S. Lipovetsky and M. Conklin. Analysis of regression in game theory approach. *Applied stochastic models in business and industry*, 17:319–330, 2001.

[22] S. Lipovetsky and M. Conklin. Meaningful regression analysis in adjusted coefficients Shapley value model. *Model Assisted Statistics and Applications*, 5: 251–264, 2010.

[23] S.C. Littlechild and G. Owen. A simple expression for the Shapley Value in a special case. *Management Science*, 20: 370–372, 1973.

[24] R. Lucchetti, P. Radrizzani, and E. Munarini. A new family of regular semivalues and applications. *International Journal of Game Theory*, 40:655–675, 2011.

[25] J.W. Mercik. Power and expectations. *Control and Cybernetics*, 26:617–621, 1997.

[26] S. Moretti, V. Fragnelli, F. Patrone, and S. Bonassi. Using coalitional games on biological networks to measure centrality and power of genes. *Bioinformatics*, 26:2721–2730, 2010.

[27] S. Moretti and F. Patrone. Transversality of the Shapley value. *TOP*, 16:1–41, 2008.

[28] S. Moretti, F. Patrone, and S. Bonassi. The class of microarray games and the relevance index for genes. *TOP*, 15:256–280, 2007.

[29] R. Myerson. Graphs and cooperation in games. *Mathematics of Operations Research*, 2:225–229, 1977.

[30] G. Owen. Multilinear extensions and the Banzhaf value. *Naval Research Logistics Quarterly*, 22:741–750, 1975.

[31] L. Sanmiquel, M. Freijó, J. Edo and J.M. Rossell. Analysis of work related accidents in the spanish mining sector from 1982-2006. *Journal of Safety Research*, 41:1–7, 2010.

[32] L.S. Shapley. A value for n-person games. In H.W. Kuhn and A.W. Tucker, editors, *Contributions to the Theory of Games II*, 307–317. Princeton University Press, Princeton, USA, 1953.

[33] A. Tsanakas and E. Desli. Measurement and pricing of risk in insurance markets. *Risk Analysis*, 25:1653–1668, 2005.

[34] F. Turnovec. Power, power indices and intuition. *Control and Cybernetics*, 26:613–615, 1997.

[35] F. Turnovec, J.W. Mercik, and M. Mazurkiewicz. Power indices methodology: decisiveness, pivots and swings. In M. Braham and F. Steffen, editors, *Power, Freedom and Voting*, 23–37. Springer Verlag, Berlin, Germany, 2008.

6 Appendix: Proofs

Proof of Proposition 2.10

First, note that:

$$a(S) = b(S) \text{ for all } S \subseteq N,\, S \ni i \quad \text{implies} \quad f_i(a) = f_i(b) \tag{5}$$

for $a, b \in \mathcal{A}^N$. In fact, by monotonicity:

$$a(S) = b(S) \text{ for all } S \subseteq N,\, S \ni i \Rightarrow a(S) \geq b(S) \text{ for all } S \subseteq N,\, S \ni i \Rightarrow f_i(a) \geq f_i(b)$$

and

$$a(S) = b(S) \text{ for all } S \subseteq N,\, S \ni i \Rightarrow a(S) \leq b(S) \text{ for all } S \subseteq N,\, S \ni i \Rightarrow f_i(a) \leq f_i(b)$$

so, $f_i(a) = f_i(b)$

- Consider the null incidence function η defined as $\eta(S) = 0$ for all $S \subseteq N$. Then by equal treatment we have that $f_i(\eta) = f_j(\eta)$ for all $i, j \in N$. By totality $\sum_{i \in N} f_i(\eta) = T(\eta) = 0$, and therefore $f_i(\eta) = 0$ for all $i \in N$.

  Let $a \in \mathcal{A}^N$. Now by (5) it follows that for any null factor $i \in N$:

  $$a(S) = 0 = \eta(S) \text{ for all } S \subseteq N,\, i \in S \quad \text{implies} \quad f_i(a) = f_i(\eta) = 0 \tag{6}$$

  Hence, we have shown that the Zero on nulls property holds.

- Write the incidence function a in its standard basis form $a = \sum_{S \subseteq N\,:\, S \neq \emptyset} a(S) a_S$. Define the index I associated to a as the number of non-zero terms in the decomposition of the incidence function a, i.e. $I = |IS|$, where $IS = \{S \subseteq N \,:\, a(S) \neq 0\} = \{S_1, S_2, ..., S_I\}$. Clearly, $0 \leq I \leq 2^n - 1$. W.l.o.g. we may consider $1 \leq I \leq 2^n - 1$ as for $I = 0$ then $IS = \emptyset$, or equivalently $a = \eta$ for which Linearity trivially holds. We prove the assertion by induction on the set IS.

  $IS = \{R\}$. Then $I = 1$ and $a = Ca_R$, i.e.,

  $$a(S) = \begin{cases} C, & \text{if } S = R \\ 0, & \text{otherwise} \end{cases}$$

  We distinguish two situations:

– $i \notin R$. We have $a(S) = 0$ for all $S \subseteq N, S \ni i$, i.e. i is a null factor, and thus by Zero on nulls property $f_i(a) = 0$. Hence,

$$0 = f_i(a) = f_i\left(\sum_{S \subseteq N\,:\,S \neq \emptyset} a(S)a_S\right)$$

On the other hand:

$$\sum_{S \subseteq N\,:\,S \neq \emptyset} a(S)f_i(a_S) = \sum_{S \subseteq N\,:\,S \neq R, S \neq \emptyset} a(S)f_i(a_S) + a(R)f_i(a_R) = \sum_{S \subseteq N\,:\,S \neq R, S \neq \emptyset} 0 \cdot f_i(a_S) + C \cdot 0$$

and Linearity holds.

– $i \in R$. For all $j \in R$ Equal treatment implies that $f_j(a) = f_i(a)$. Combined with totality this implies that $f_i(a) = \dfrac{C}{|R|}$, but also $f_i(a_R) = \dfrac{1}{|R|}$. Thus,

$$\frac{C}{|R|} = f_i(a) = f_i\left(\sum_{S \subseteq N\,:\,S \neq \emptyset} a(S)a_S\right)$$

On the other hand:

$$\begin{aligned}\sum_{S \subseteq N\,:\,S \neq \emptyset} a(S)f_i(a_S) &= \sum_{S \subseteq N\,:\,S \neq R, S \neq \emptyset} a(S)f_i(a_S) \;+\; a(R)f_i(a_R) \\ &= \sum_{S \subseteq N\,:\,S \neq R, S \neq \emptyset} 0 \cdot f_i(a_S) \;+\; C \cdot \frac{1}{|R|} = \frac{C}{|R|}\end{aligned}$$

and Linearity holds.

Assume the assertion holds for I; prove it holds for $I = I + 1$.
Assume that Linearity holds for any $b \in \mathcal{A}^N \;:\; b = \sum_{k=1}^{I} b(S_k)a_{S_k}$. Let $a \in \mathcal{A}^N \;:\; a = \sum_{k=1}^{I+1} a(S_k)a_{S_k}$, i.e. the other elements of $a(S)$ are zero, and let $IS^+ = \{S_1, S_2, ..., S_{I+1}\}$. Let $R = \cap_{k=1}^{I+1} S_k$.
Suppose that $i \notin R$.
Consider the incidence function

$$\bar{a} = \sum_{S_k \in IS^+\,:\,S_k \ni i} a(S_k)a_{S_k}$$

Then $\bar{a}$ has an index of at most I with regard to $i \notin R$. Furthermore,

$$a(S) = \bar{a}(S) \quad \text{for all } S \subseteq N,\, i \in S.$$

Thus by induction hypothesis and monotonicity, it can be concluded that

$$\begin{aligned}f_i(a) &= f_i(\bar{a}) = \sum_{S_k \in IS^+\,:\,S_k \ni i} a(S_k)f_i(a_{S_k}) \\ &= \sum_{S_k \in IS^+\,:\,S_k \ni i} a(S_k)f_i(a_{S_k}) \;+\; \sum_{S_k \in IS^+\,:\,S_k \not\ni i} a(S_k)f_i(a_{S_k}) \\ &= \textstyle\sum_{k=1}^{I+1} a(S_k)f_i(a_{S_k})\end{aligned}$$

where the third equality holds because $f_i(a_{S_k}) = 0, S_k \in IS^+ \;:\; S_k \not\ni i$.

Next suppose that $i \in R$.
By Equal treatment, $f_i(a)$ is a constant c for all factors in R. By equal treatment and totality $f_i(a_S) = 1/|S|$ for all $S \subseteq N$, $i \in S$.

By totality and because $a(S) = 0$ if $R \not\subseteq S$:

$$\sum_{i \in R} f_i(a) + \sum_{i \notin R} f_i(a) = \sum_{S \subseteq N} a(S) = \sum_{R \subseteq S \subseteq N} a(S)$$

Thus,

$$c|R| + \sum_{R \subseteq S \subseteq N} \frac{|S| - |R|}{|S|} a(S) + \sum_{R \subseteq S \subseteq N} a(S)$$

Hence,

$$f_i(a) = c = \sum_{R \subseteq S \subseteq N} \frac{a(S)}{|S|} = \sum_{S \subseteq N, i \in S} a(S) \frac{1}{|S|} = \sum_{S \subseteq N, i \in S} a(S) f_i(a_S)$$

□

Proof of Proposition 3.1

Existence: It is obvious that $\mathfrak{F}$ satisfies the four stated properties.
Uniqueness: For any set S ($S \neq \emptyset$) of factors let us consider the basic incidence function a_S that counts one accident for only the set of factors S:

$$a_S(R) = \begin{cases} 0 & \text{if } R \neq S \\ \\ 1 & \text{if } R = S \end{cases}$$

Now each a can be written as:

$$a = \sum_{S \subseteq N : S \neq \emptyset} a(S) \cdot a_S$$

Taking into account totality, zero on nulls and equal treatment it follows that the value of f for a_S is

$$f_i(a_S) = \begin{cases} \dfrac{1}{|S|} & \text{if } i \in S \\ \\ 0 & \text{if } i \notin S \end{cases}$$

and therefore, from linearity property it follows that

$$f_i(a) = \sum_{S \subseteq N : i \in S} \frac{a(S)}{|S|}$$

Thus, $f_i(a) = \mathfrak{F}_i(a)$ for all $i \in N$ and $a \in \mathcal{A}_N$. Hence, $f = \mathfrak{F}$. □

Proof of Proposition 3.5

To see that the game v_a is convex, first note that the value of $v_a(S \cup T) + v_a(S \cap T)$ can be expressed by:

$$\sum_{R \subseteq S \cup T} a(R) + \sum_{R \subseteq S \cap T} a(R) = \sum_{R \subseteq S} a(R) + \sum_{\substack{R \subseteq T : \\ R \cap (T - S) \neq \emptyset}} a(R) + \sum_{\substack{R \subseteq S \cup T : \\ S \Delta T \subseteq R}} a(R) + \sum_{R \subseteq S \cap T} a(R)$$

where $S \Delta T = (S - T) \cup (T - S)$ is the symmetric difference. On the other hand, the expression $v_a(S) + v_a(T)$ can be written:

$$\sum_{R \subseteq S} a(R) + \sum_{R \subseteq T} a(R) = \sum_{R \subseteq S} a(R) + \sum_{R \subseteq S \cap T} a(R) + \sum_{\substack{R \subseteq T : \\ R \cap (T - S) \neq \emptyset}} a(R)$$

Thus, the expression $v_a(S \cup T) + v_a(S \cap T) \geq v_a(S) + v_a(T)$ can be simplified to:

$$\sum_{\substack{R \subseteq S \cup T : \\ S \Delta T \subseteq R}} a(R) \geq 0,$$

that holds by the non-negativity of the incidence function a. □

Proof of Proposition 3.6 First note that

$$v_a(S \cup \{i\}) - v_a(S) = \sum_{T \subseteq S} a(T \cup \{i\}) \qquad \text{for all } \; S \subseteq N \setminus \{i\}$$

Thus,

$$\Phi_i[v_a] = \sum_{S \subseteq N \setminus \{i\}} \left[\sum_{T \subseteq N \setminus \{i\} \, : \, S \subseteq T} \gamma_n(t) \right] a(S \cup \{i\})$$

where $|T| = t$ and $\gamma_n(t) = 1/[n\binom{n-1}{t}]$. Let $\Gamma_n(s)$ be the coefficient of $a(S \cup \{i\})$ in the latter expression, i.e.,

$$\Gamma_n(s) = \sum_{T \subseteq N \setminus \{i\} \, : \, S \subseteq T} \gamma_n(t) \qquad \text{for all } \; S \subseteq N \setminus \{i\}$$

To see that $\Phi_i[v_a]$ coincides with $\mathfrak{F}_i[a]$ we need to prove that $\Gamma_n(s) = 1/(s+1)$ for all $s = 0, 1, \ldots, n-1$. We proceed by induction on s.

Assume $s = 0$. Then $\Gamma_n(0) = \sum_{k=0}^{n-1} \binom{n-1}{k} \gamma_n(k) = \sum_{k=0}^{n-1} 1/n = 1$.

Assume that $\Gamma_n(s) = 1/(s+1)$, we need to prove that $\Gamma_n(s+1) = 1/(s+2)$. We express $\Gamma_n(s+1)$ as a function of $\Gamma_n(s)$ in the following way:

$$\Gamma_n(s+1) = \Gamma_n(s) + B(s) - \gamma_n(s)$$

where

$$\begin{aligned} B(s) &= \sum_{k=1}^{n-2-s} \left(\tbinom{n-2-s}{k-1} - \tbinom{n-1-s}{k} \right) \gamma(s+k) \\ &= \frac{(n-2-s)!}{n!} \sum_{k=1}^{n-2-s} \left(\frac{(s+k)!(k-n+1+s)}{k!} \right) \end{aligned}$$

has a known sum, so that:

$$B(s) - \gamma_n(s) = -\frac{1}{(s+1)(s+2)}$$

and by induction hypothesis, i.e., $\Gamma_n(s+1) = 1/(s+1) - 1/((s+1)(s+2)) = 1/(s+2)$. □